Multi-subject MEG/EEG source imaging with sparse multi-task regression

Hicham Janati[a,c,*], Thomas Bazeille[a], Bertrand Thirion[a], Marco Cuturi[b,c], Alexandre Gramfort[a]

[a] *Université Paris-Saclay, Inria, CEA, France*
[b] *Google, France*
[c] *ENSAE / CREST, France*

Abstract

Magnetoencephalography and electroencephalography (M/EEG) are non-invasive modalities that measure the weak electromagnetic fields generated by neural activity. Estimating the location and magnitude of the current sources that generated these electromagnetic fields is a challenging ill-posed regression problem known as *source imaging*. When considering a group study, a common approach consists in carrying out the regression tasks independently for each subject. An alternative is to jointly localize sources for all subjects taken together, while enforcing some similarity between them. By pooling all measurements in a single multi-task regression, one makes the problem better posed, offering the ability to identify more sources and with greater precision. The Minimum Wasserstein Estimates (MWE) promotes focal activations that do not perfectly overlap for all subjects, thanks to a regularizer based on Optimal Transport (OT) metrics. MWE promotes spatial proximity on the cortical mantel while coping with the varying noise levels across subjects. On realistic simulations, MWE decreases the localization error by up to 4 mm per source compared to individual solutions. Experiments on the Cam-CAN dataset show a considerable improvement in spatial specificity in population imaging. Our analysis of a multimodal dataset shows how multi-subject source localization closes the gap between MEG and fMRI for brain mapping.

Keywords: Brain, Inverse modeling, EEG / MEG source imaging

1. Introduction

Magnetoencephalography (MEG) measures the magnetic field surrounding the head, while Electroencephalography (EEG) measures the electric potential at the surface of the scalp. Both can do so with a temporal resolution of less than a millisecond [4, 27]. Localizing the

*Corresponding author.
Email addresses: hicham.janati@inria.fr (Hicham Janati), thomas.bazeille@inria.fr (Thomas Bazeille), bertrand.thirion@inria.fr (Bertrand Thirion), cuturi@google.com (Marco Cuturi), alexandre.gramfort@inria.fr (Alexandre Gramfort)

October 14, 2019

underlying neural activity at the origin of the signals is a linear inverse problem known as *source imaging* [5, 6, 44, 67]. From a statistical perspective, it can be regarded as a *linear regression problem* in high dimension. This linearity follows directly from Maxwell equations. However, this inverse problem is inherently "ill-posed": Indeed, the number of potential sources is larger than the number of MEG and EEG sensors, which implies that, even in the absence of noise, different neural activity patterns could result in the same electromagnetic field measurements. This fact makes M/EEG source imaging particularly challenging in the presence of multiple simultaneous active regions in the brain.

To limit the set of possible solutions, prior hypotheses on the nature of the source distributions are necessary. The minimum-norm estimates (MNE) for instance are based on ℓ_2 Tikhonov regularization which leads to a linear solution [25]. An ℓ_1 norm penalty was also proposed by Uutela et al. [63], modeling the underlying neural pattern as a sparse collection of focal dipolar sources, hence their name "Minimum Current Estimates" (MCE). These methods have inspired a series of contributions in source localization techniques relying on noise normalization such as dSPM [11] and sLORETA [11, 52] to correct for the depth bias [2] or block-sparse norms such as MxNE [56] and TF-MxNE [24] to leverage the spatio-temporal dynamics of MEG signals. If other imaging data are available such as fMRI [50, 70] or diffusion MRI [12], they can be used as prior information for example in hierarchical Bayesian models [55]. While such techniques have had some success, source estimation in the presence of complex multi-dipole configurations remains a challenge. To address it, one idea is to leverage the anatomical and functional diversity of multi-subject datasets to improve localization results.

This idea of using multi-subject information to improve the spatial accuracy of M/EEG source imaging has been proposed before in the neuroimaging literature. Larson et al. [38] hypothesized that different anatomies across subjects allow for point spread functions that agree on a main activation source but differ elsewhere. Averaging across subjects thereby increases the accuracy of source localization. On fMRI data, Varoquaux et al. [64] proposed a probabilistic dictionary learning model to infer activation maps jointly across a cohort of subjects. A similar idea lead Litvak and Friston [41] to propose a Bayesian hierarchical model to cope with inter-subject functional variability. Their model however relied on a common source space for all subjects. This assumption was relaxed by Kozunov and Ossadtchi [36] who proposed a similar Bayesian model. However, their formulation involves a set of large $(p \times p)$ covariance matrices.

Source imaging for a set of subjects can be formulated as solving a set of coupled regression problems. In the statistical machine learning literature such supervised learning problems are commonly referred to as *multi-task* prediction problems. In the M/EEG source imaging literature, to our knowledge, the only contribution formulating the problem as a multi-task regression model employs a Group Lasso with an ℓ_{21} block sparse norm [39]. Yet this work forces every potential neural source to be either active for all subjects, or for none of them.

The assumption of perfectly ovelapping functional activity across subjects gets even more unrealistic as we aim for fine spatial resolution in the order of millimeters. In this work, one investigates several multi-task regression models that relax this assumption. One of them is the multi-task Wasserstein (MTW) model [32]. MTW is defined through an

Unbalanced Optimal Transport (UOT) metric that promotes support proximity across regression coefficients. However, applying MTW to group level data assumes that the signal-to-noise ratio is the same for all subjects. To alleviate this problem, one can infer both the sources and the noise variance for each subject and scale the regularization according to the level of noise. Following similar ideas that lead to the concomitant Lasso [46, 51, 58] or the multi-task Lasso [43], the Minimum Wasserstein Estimates (MWE) was first proposed in [31]. However both MTW and MWE rely on convex ℓ_1 norm penalties which tend to promote sparse solutions at the expense of an amplitude bias.

One way to mitigate amplitude bias of convex regularizations is to make use of non-convex ones, such as ℓ_p pseudo norms with $0 < p < 1$. Due to their non-convexity, these pseudo norms provide better proxies for the ℓ_0 norm and can thereby promote more sparsity with no amplitude bias [19, 56]. Gasso et al. [18] studied a broad family of non-convex penalties and showed that the $\ell_{0.5}$ regularized regression is equivalent to a type of what was previously known as re-weighted or adaptive Lasso [8].

The paper is organized as follows. In Section 2, one explains how multi-subject M/EEG source imaging can be cast as a multi-task regression problem. Methods from the literature that adopt this approach are briefly recalled. Next some background on UOT metrics are presented. Then the reweighted minimum Wasserstein estimates ($\text{MWE}_{0.5}$) method is presented. UOT is proposed to cope with inter-subject spatial variability, $\ell_{0.5}$ pseudo-norm is used to limit amplitude estimation bias and concomitant estimation is exploited to handle subjects affected by different noise levels. Finally experimental results on both simulations and two public MEG datasets [60, 66] are reported.

A preliminary version of this work was presented at the international conference on Information Processing in Medical Imaging (IPMI) [31].

Notation. We denote by $\mathbb{1}_p$ the vector of ones in $\mathbb{R}^p$ and by I_n the square identity matrix of dimension n. $[\![q]\!]$ denotes the set $\{1, \dots, q\}$ for any integer $q \in \mathbb{N}$. The set of vectors in $\mathbb{R}^p$ with non-negative (resp. positive) entries is denoted by $\mathbb{R}^p_+$ (resp. $\mathbb{R}^p_{++}$). On matrices, log, exp and the division operator are applied elementwise. We use $\odot$ for the elementwise multiplication between matrices or vectors. If $\mathbf{X}$ is a matrix, $\mathbf{X}_{i\cdot}$ denotes its i^{th} row and $\mathbf{X}_{\cdot j}$ its j^{th} column. The scalar product between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$. We define the Kullback-Leibler (KL) divergence between two positive vectors by $\text{KL}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \log(\mathbf{x}/\mathbf{y}) \rangle + \langle \mathbf{y} - \mathbf{x}, \mathbb{1}_p \rangle$ with the continuous extensions $0 \log(0/0) = 0$ and $0 \log(0) = 0$. We also make the convention $\mathbf{x} \neq 0 \Rightarrow \text{KL}(\mathbf{x}|0) = +\infty$. The entropy of $\mathbf{x} \in \mathbb{R}^n$ is defined as $H(\mathbf{x}) = -\langle \mathbf{x}, \log(\mathbf{x}) - \mathbb{1}_p \rangle$. The same definition applies for matrices with an element-wise double sum. The ℓ_{21} norm of a matrix $A \in \mathbb{R}^{p \times S}$ is defined as $\sum_{i=1}^{p} \|A_{i\cdot}\|_2$.

2. Source imaging as a multi-task regression problem

In this section, one shows how the M/EEG inverse problem at the group level can be cast as a multi-task regression problem to estimate jointly source locations and magnitudes for multiple subjects.

Cortically constrained source modeling. Using a segmentation of the MRI scan of each subject, the positions of potential sources are constructed as a set of coordinates uniformly distributed on the cortical surface [11]. Since synchronized currents flowing along the apical dendrites of cortical pyramidal neurons are thought to be mostly responsible for M/EEG signals [48], we constrain the dipole orientations to be normal to the cortical surface. Thus, we model the current density as a set of focal current dipoles with fixed positions and orientations. The purpose of source localization is then to infer their amplitudes. The ensemble of possible candidate dipoles forms the *source space*.

Forward modeling. Let n denote the number of sensors (EEG and/or MEG) and p the number of sources. Following Maxwell's equations, at each time instant, the electromagnetic fields $\mathbf{b} \in \mathbb{R}^n$ are a linear combination of the current density $\mathbf{x} \in \mathbb{R}^p : \mathbf{b} = \mathbf{L}\mathbf{x}$. The linear forward operator $\mathbf{L} \in \mathbb{R}^{n \times p}$ is called the *leadfield* or *gain matrix*. However, one observes noisy measurements $\mathbf{y} \in \mathbb{R}^n$ given by:

$$\mathbf{y} = \mathbf{b} + \varepsilon = \mathbf{L}\mathbf{x} + \varepsilon \ , \tag{1}$$

where ε is the noise vector that can be assumed Gaussian distributed $\mathcal{N}(0, \Sigma)$. In practice, $\mathbf{L}$ is computed by solving Maxwell's equations using for example a Boundary element method [26, 37, 45].

Whitening. Since the signals of MEG sensors are correlated by design, the noise covariance matrix Σ is not diagonal. For the inverse problem to be cast as a least squares problem, we perform a whitening transformation of the data. We estimate Σ during the baseline (before stimulus) using a cross-validation estimator provided in the MNE software [14, 21, 22]. Given a noise covariance matrix estimate $\hat{\Sigma}$, the whitening step amounts to computing the transformed data $\hat{\Sigma}^{-\frac{1}{2}}\mathbf{y}$ and $\hat{\Sigma}^{-\frac{1}{2}}\mathbf{L}$. In the rest of this paper, we assume that the data are whitened.

Source localization. Source localization consists in solving in $\mathbf{x}$ the inverse problem (1) which can be cast as a least squares problem:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 \ . \tag{2}$$

Since $n \ll p$, problem (2) is ill-posed and additional constraints on the solution $\mathbf{x}^\star$ are necessary. These constraints materialize the type of prior knowledge one has on the source estimates and are commonly applied through ℓ_p norms [28, 49, 56]. To favor weak distributed currents over focal sources, one can use a squared ℓ_2 regularization, leading to minimum-norm Estimates (MNE) [25]:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \ , \tag{3}$$

where $\lambda > 0$ is a tuning hyperparameter. Both dSPM [11] and sLORETA [52] are variants of MNE that tackle noise normalization.

When analyzing evoked responses, one can instead promote source configurations made of a few focal sources, e.g. using the ℓ_1 norm. This regularization leads to problem (4) called minimum current estimates (MCE) [63], also known in the machine learning community as the Lasso [61].

$$\mathbf{x}^\star = \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{L}\mathbf{x}\|_2^2+\lambda\|\mathbf{x}\|_1 \ , \tag{4}$$

Depth weighting. The inverse solutions discussed above are inherently biased towards sources in the superficial layers of the cortex [40]. Indeed, deep sources require larger amplitude values than superficial ones to produce a magnetic field with similar strength. To circumvent this problem, we normalize the columns of the leadfield $\mathbf{L}$ by a fraction of their norms. In all our experiments we use a depth weighting of 0.9. Formally, every column $\mathbf{L}_{.j}$ is normalized by $\|\mathbf{L}_{.j}\|^{0.9}$ [24, 40].

Common source space. Here one proposes to go beyond the classical pipeline and carry out source localization jointly for S subjects. To do so, dipole positions (features) must correspond spatially across subjects. Using an anatomical alignment procedure, the source space of each subject is mapped from an average brain. This process one calls *morphing* uses the sulci and gyri patterns which are matched in an auxiliary spherical inflated cortical surface [16]. The resulting leadfields $\mathbf{L}^{(1)}, \dots, \mathbf{L}^{(S)}$ have therefore the same shape $(n \times p)$ with aligned columns; a given column maps to the same brain region across all subjects.

Multi-task framework. Jointly estimating the current density $\mathbf{x}^{(s)}$ of each subject s can be expressed as a multi-task regression problem where some coupling prior is assumed on $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}$ through a penalty Ω:

$$\min_{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(S)}\in\mathbb{R}^p} \frac{1}{2n}\sum_{s=1}^{S}\|\mathbf{y}^{(s)}-\mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2+\Omega(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(S)}) \ . \tag{5}$$

The work of Lim et al. [39] embraces this multi-task framework where the joint penalty is defined through an ℓ_{21} mixed norm [20]:

$$\Omega(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(S)}) = \sum_{j=1}^{p}\sqrt{\sum_{s=1}^{S}{\mathbf{x}_j^{(s)}}^2} \ . \tag{6}$$

Following the work of [32], one can define Ω using an UOT metric. By doing so, one does not require an exact spatial correspondence between active sources in the group of subjects as enforced by [39, 41].

3. Optimal transport background

Optimal transport theory offers the mathematical tools to compare probabilities embedded in a metric space [65]. In a discrete space, probability measures can be represented as

histograms, i.e a finite dimensional vector of weights. Each entry in such a vector is the mass of a bin in the histogram. In the present context, a bin corresponds to a vertex of the source space, and the weight quantifies the amplitude of the neural activation at this location. In this case, the metric required by optimal transport boils down to the definition of a distance between the bins of the histogram. It is known as the *ground metric*. For the application considered here, these tools will enable us to define a notion of similarity between source estimates, while taking into account the folded geometry of the cortical mantel.

We now provide some background material on unbalanced OT (UOT), which allows to consider distances between non-normalized histograms. Entries of such histograms do not sum to one, and therefore do not define probabilities. Consider the space (E, d) where each element of $E = \{1, \dots, p\}$ corresponds to a vertex of the cortical source space. Let $\mathbf{M}$ be the $p \times p$ matrix where the entry $\mathbf{M}_{ij}$ is the geodesic distance between vertices i and j. It encodes the ground metric. Kantorovic and Rubinshteïn [33] initially defined an optimal transport distance for normalized histograms (probability measures) on E. However, it can easily be extended to non-normalized histograms [9, 17].

Distances between non-normalized histograms. Let $\mathbf{a}, \mathbf{b}$ be two normalized histograms on E. Assuming that transporting a fraction of mass $\mathbf{P}_{ij}$ from i to j is given by $\mathbf{P}_{ij}\mathbf{M}_{ij}$, the total cost of transport is given by $\langle \mathbf{P}, \mathbf{M} \rangle = \sum_{ij} \mathbf{P}_{ij}\mathbf{M}_{ij}$. To compare a and b one is interested by the minimum amount of mass one needs to move to transport a to b. Minimizing this total cost with respect to $\mathbf{P}$ must be carried out on the set of feasible transport plans with marginals $\mathbf{a}$ and $\mathbf{b}$. It amounts to enforcing that no mass appeared or disappeared during the transport. The (normalized) Wasserstein distance reads:

$$\mathrm{W}(\mathbf{a}, \mathbf{b}) = \min_{\substack{\mathbf{P} \in \mathbb{R}_+^{p \times p} \\ \mathbf{P}\mathbb{1} = \mathbf{a}, \mathbf{P}^\top \mathbb{1} = \mathbf{b}}} \langle \mathbf{P}, \mathbf{M} \rangle \ . \tag{7}$$

In practice, if $\mathbf{a}$ and $\mathbf{b}$ are current densities formed each by one focal active source of amplitude one (normalized), $\mathrm{W}(\mathbf{a}, \mathbf{b})$ will quantify the geodesic distance between the two locations along the curved geometry of the cortex. The distance W is however also adapted to multi-dipoles source configurations and possibly spatially extended sources. It is also known as the Earth mover distance (EMD). This property, and the connection with the notion of mass displacement, makes OT metrics adequate for assessing the proximity of functional activations across subjects. We illustrate this behavior of the Wasserstein distance in Figure 1. We compute the Wasserstein distances between a fixed brain activation $\mathbf{a}$ spread over 4 vertices (green) and a set of activations $\mathbf{b}$ concentrated on one vertex. The color of each vertex denotes the Wasserstein distance between $\mathbf{a}$ and $\mathbf{b}$. If $\mathbf{a}$ was restricted to only 1 vertex, the Wasserstein distance would be equal to the geodesic distance between the pair of activation foci of $\mathbf{a}$ and $\mathbf{b}$. This distance can be seen as a generalization of the geodesic that compares neural patterns that are spread over multiple vertices of the cortical mesh.

To allow $\mathbf{a}, \mathbf{b}$ to be non-normalized, the marginal constraints in (7) can be relaxed using a KL divergence:

$$\min_{\mathbf{P} \in {\mathbb{R}_+}^{p \times p}} \langle \mathbf{P}, \mathbf{M} \rangle + \gamma \mathrm{KL}(\mathbf{P}\mathbb{1} | \mathbf{a}) + \gamma \mathrm{KL}(\mathbf{P}^\top \mathbb{1} | \mathbf{b}) \ , \tag{8}$$

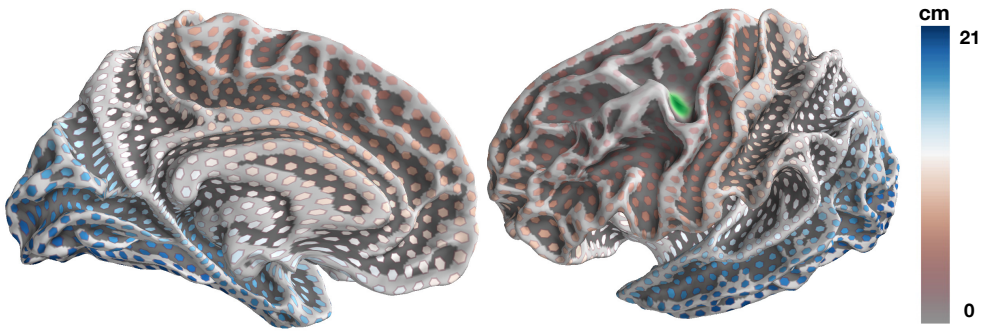

Fig. 1. Levels of the W distance in cm (a.k.a Earth mover distance) computed between a fixed blurred activation $\mathbf{a}$ (green) and all focal activations $\mathbf{b}$ of the triangular mesh of the cortex. **Left:** Medial view. **Right:** Lateral view.

where $\gamma > 0$ is a hyperparameter that enforces a fit to the marginals.

Entropy regularization. Entropy regularization was introduced by Cuturi [10] to propose a faster and more robust alternative to the direct resolution of the linear programming problem (7). Formally, this amounts to minimizing the loss $\langle \mathbf{P}, \mathbf{M} \rangle - \varepsilon H(\mathbf{P})$ where $\varepsilon > 0$ is a tuning hyperparameter. This penalized loss function can be written: $\varepsilon \mathrm{KL}(\mathbf{P}, e^{-\frac{\mathbf{M}}{\varepsilon}})$ up to a constant [7]. Combining entropy regularization with marginal relaxation in (8), we get the unbalanced Wasserstein distance W_u as introduced independently by Chizat et al. [9], Frogner et al. [17]:

$$W_u(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbb{R}_+^{p \times p}} \varepsilon \mathrm{KL}(\mathbf{P} | e^{-\frac{\mathbf{M}}{\varepsilon}}) + \gamma \mathrm{KL}(\mathbf{P}\mathbb{1} | \mathbf{a}) + \gamma \mathrm{KL}(\mathbf{P}^\top \mathbb{1} | \mathbf{b}) \ . \tag{9}$$

Generalized Sinkhorn. Problem (9) can be solved as follows. Let $\mathbf{K} = e^{-\frac{\mathbf{M}}{\varepsilon}}$ and $\psi = \gamma/(\gamma + \epsilon)$. Starting from two vectors $\mathbf{u}, \mathbf{v}$ set to $\mathbb{1}$ and iterating the scaling operations $\mathbf{u} \leftarrow (\mathbf{a}/\mathbf{K}\mathbf{v})^\psi$, $\mathbf{v} \leftarrow \left(\mathbf{b}/\mathbf{K}^\top \mathbf{u}\right)^\psi$ until convergence, the minimizer of (9) can be computed as $P^\star = (\mathbf{u}_i \mathbf{K}_{ij} \mathbf{v}_j)_{i,j \in [\![p]\!]}$. This algorithm is a generalization of the Sinkhorn algorithm [35]. Since it involves matrix-matrix operations, it benefits from parallel hardware, such as GPUs.

Extension to $\mathbb{R}^p$. Finally, as a M/EEG source estimates can be positive or negative, we extend the Wasserstein distance W_u to signed measures. We adopt a similar idea to what was suggested in [32, 42, 53] using a decomposition into positive and negative parts, $\mathbf{a} = \mathbf{a}^+ - \mathbf{a}^-$ where $\mathbf{a}^+ = \max(\mathbf{a}, 0)$ and $\mathbf{a}^- = \max(-\mathbf{a}, 0)$. For any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, we define the generalized Wasserstein distance as:

$$\widetilde{W}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} W_u(\mathbf{a}^+, \mathbf{b}^+) + W_u(\mathbf{a}^-, \mathbf{b}^-) \ . \tag{10}$$

Note that $W_u(\mathbf{0}, \mathbf{0}) = 0$ (both optimal transport plans must be $\mathbf{0}$), thus on positive measures $\widetilde{W} = W_u$. For the sake of convenience, we refer to $\widetilde{W}$ in (10) as the Wasserstein *distance*, even though $\widetilde{W}(\mathbf{a}, \mathbf{a}) \neq 0$ in general. In practice, this extension allows to compare source

densities across subjects taking into account their polarity. Using common conventions in M/EEG source imaging, positive currents are flowing out of the cortex (from deep cortical layers to superficial ones), while negative currents are flowing into the cortex.

Wasserstein barycenters. As introduced by Agueh and Carlier [1], the Wasserstein barycenter $\bar{\boldsymbol{x}}$ of a set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_S$ is defined as the Fréchet mean of $\widetilde{W}$ across the inputs [1]:

$$\bar{\boldsymbol{x}} = \operatorname*{arg\,min}_{\mathbf{x} \in \mathbb{R}^p} \frac{1}{S} \sum_{s=1}^{S} \widetilde{W}(\mathbf{x}^{(s)}, \mathbf{x}) \ , \tag{11}$$

In practice, one can write $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}^{(s)+} - \bar{\boldsymbol{x}}^{(s)-}$. Since the positive and negative parts in (10) are not coupled, computing $\bar{\boldsymbol{x}}$ only requires a slightly modified version of Generalized Sinkhorn (Algorithm 1) for each part.

Intuitively, one can think of the Wasserstein barycenter as a "spatial" averaging technique. Figure 2 illustrates this intuition: the Wasserstein barycenter of the three green activations is located right in their middle location. This idea of averaging source estimates across multiple subjects using Optimal transport was previously proposed by Gramfort et al. [23]. Their method – even though based on a different extension of the Earth Mover distance W – has shown considerable lower smoothing compared to Euclidean averaging and better identification of focal activations. However, it is only carried out as a post-processing step. Here we argue that including optimal transport in the inverse solver upfront allows to improve both the inference of the individual inverse solution of each subject and the average activation across subjects.

Algorithm 1 Generalized Sinkhorn - barycenter computation [9]

Input: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)} \in \mathbb{R}_+^p$
Output: Wasserstein barycenter ($\bar{\boldsymbol{x}}$) of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$ and marginals $\mathbf{m}^1, \ldots, \mathbf{m}^{(S)}$.
Initialize for $(s = 1, \ldots, S)$ $(u^{(s)}, v^{(s)}) = (\mathbb{1}, \mathbb{1})$,
repeat
  for $s = 1$ **to** S **do**
    $u^{(s)} \leftarrow \left(\mathbf{x}^{(s)} / K v^{(s)}\right)^{\psi}$
  end for
  $\bar{\boldsymbol{x}} \leftarrow \left(\frac{1}{S} \sum_{s=1}^{S} (v^{(s)} \odot K^\top u^{(s)})^{1-\psi}\right)^{\frac{1}{1-\psi}}$
  for $s = 1$ **to** S **do**
    $v^{(s)} \leftarrow \left(\bar{\boldsymbol{x}} / K^\top u^{(s)}\right)^{\psi}$
  end for
until convergence
for $t = 1$ **to** T **do**
  $\mathbf{m}^{(s)} = u^{(s)} \odot K v^{(s)}$
end for

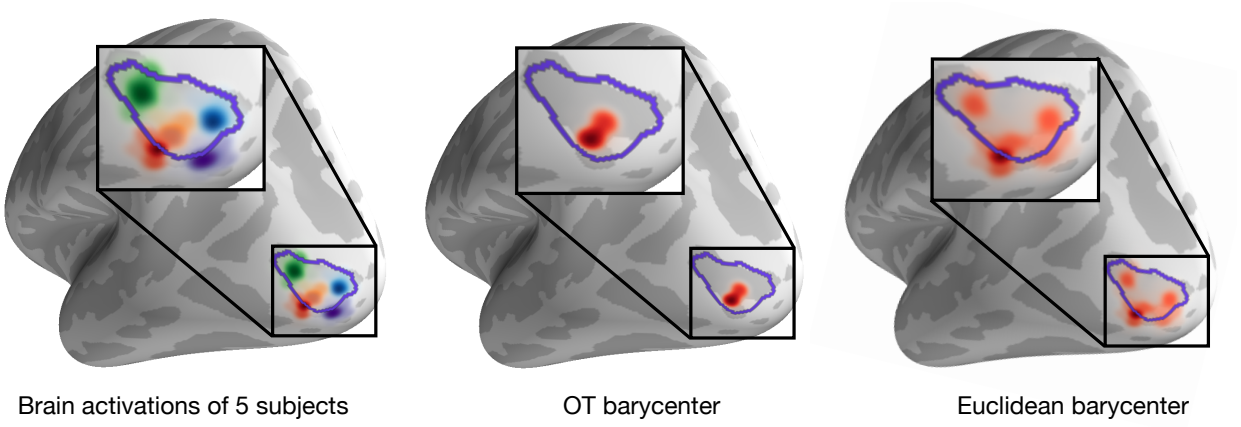

Fig. 2. Illustration of the Wasserstein barycenter $\bar{\boldsymbol{x}}$ (middle) of 5 activations inputs $\mathbf{x}^{(s)}$ (left) with random amplitudes between 20 and 30 nAm in the *middle and occipital lunatus sulcus* defined by the *aparc.a2009s* segmentation. $\bar{\boldsymbol{x}}$ is located at the average location of the inputs with an average amplitude levels. The Euclidean barycenter (right) is the usual mean: it creates undesirable blurring.

4. Minimum Wasserstein Estimates

The MTW$_q$ model. The multi-task Wasserstein model of order q (MTW$_q$), with $0 < q \leq 1$, is the specific case of (5) with a penalty Ω promoting both sparsity and spatial proximity between activation foci. The regularization term reads:

$$\Omega_{\mathrm{MTW}_q}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \stackrel{\text{def}}{=} \mu \min_{\bar{\boldsymbol{x}} \in \mathbb{R}^p} \frac{1}{S} \sum_{s=1}^{S} \widetilde{W}(\mathbf{x}^{(s)}, \bar{\boldsymbol{x}}) + \lambda \|\mathbf{x}^{(s)}\|_q \ , \tag{12}$$

where $\mu, \lambda \geq 0$ are tuning hyperparameters. The minimized OT sum in (12) measures the average distance between all the $\mathbf{x}^{(s)}$ and their Wasserstein barycenter $\bar{\boldsymbol{x}}$. It can thus be seen as quantification of the spatial variability of the source estimates. If $q = 1$, one falls back to the MTW model of Janati et al. [32].

Minimum Wasserstein Estimates. One of the drawbacks of MTW is that λ is common to all subjects. Indeed, the loss considered in MTW implicitly assumes that the level of noise is the same across subjects. Following the work of [46] on the smoothed concomitant Lasso, we propose to extend MTW by inferring the specific noise standard deviation $\sigma^{(s)}$ along with the regression coefficient $\mathbf{x}^{(s)}$ of each subject. This allows to scale the weight of the ℓ_q regularization according to the level of noise. The Minimum Wasserstein Estimates (MWE$_q$) model reads:

$$\min_{\substack{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)} \in \mathbb{R}^p \\ \sigma^{(1)}, \dots, \sigma^{(S)} \in [\sigma_0, +\infty]}} \sum_{s=1}^{S} \frac{1}{2n\sigma^{(s)}} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)} \mathbf{x}^{(s)}\|_2^2 + \frac{\sigma^{(s)}}{2} + \Omega_{\mathrm{MTW}_q}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \ , \tag{13}$$

where σ_0 is a pre-defined constant. This lower bound constraint prevents from a null standard deviation and divisions by zero, while also making the feasible set a convex domain. In practice σ_0 can be set for example using prior knowledge on the variance of the data or as a small fraction of the initial estimate of the standard deviation $\sigma_0 = \alpha \min_s \frac{\|\mathbf{y}^{(s)}\|}{\sqrt{n}}$. In practice we adopt the second option and set $\alpha = 0.01$, although we make sure that it does not affect the solutions by checking that that the estimated $\hat{\sigma}^{(s)}$ are stricly superior to σ_0.

Reweighted Minimum Wasserstein Estimates. Several studies have shown that non-convex ℓ_q with $0 < q < 1$ not only reduce the amplitude bias but also promote a more accurate support estimation [8, 18, 57]. We define the reweighted Minimum Wasserstein estimates as MWE_q with $q = 0.5$. The resulting optimization problem can be solved in a sequence of weighted instances of MWE_1. This reweighting scheme can be seen as a majorization-minimization algorithm [56]. Indeed, since the $\ell_{0.5}$ pseudo-norm is separable and concave, it can be upper bounded by its element-wise derivative. The reweighting amounts to iteratively minimizing instances of MWE_1 with weighted ℓ_1 norms and updating the upper bound. These steps are summarized in Algorithm 2. When some $\mathbf{x}_j^{(s)} = 0$, the majorization step will cause an overflow error $\mathbf{w}_j^{(s)} = +\infty$ which corresponds to an infinite amount of regularization and thus leads to $\mathbf{x}_j^{(s)} = 0$. In practice, one can simply filter out the corresponding features or set $\mathbf{w}_j^{(s)} = \frac{1}{2\sqrt{|\mathbf{x}_j^{(s)}| + \eta}}$ where η is a small value as proposed by Gasso et al. [18]. We adopt this strategy and set $\eta = 10^{-6}$.

Algorithm for the MWE_1 subproblems. We can now explain how to solve the MWE_1 subproblems. By combining (9), (10) and (13), we obtain an objective function taking as arguments $\big((\mathbf{x}^{(s)+})_s, (\mathbf{x}^{(s)-})_s, (\mathbf{P}^{(s)+})_s, (\mathbf{P}^{(s)-})_s, \bar{\boldsymbol{x}}^+, \bar{\boldsymbol{x}}^-, (\sigma^{(s)})_s\big)$. This function restricted to all parameters except $(\sigma^{(s)})_s$ is jointly convex [32]. Moreover, each $\sigma^{(s)}$ is only coupled with the variable $\mathbf{x}^{(s)}$. The restriction on every pair $(\mathbf{x}^{(s)}, \sigma^{(s)})$ is also jointly convex [46]. Thus the problem is jointly convex in all its variables. It can be minimized by alternating optimization. To justify the convergence of such an algorithm, one needs to notice that the non-smooth ℓ_1 norms in the objective are separable [62]. The update with respect to each $\sigma^{(s)}$ is given by solving the first order optimality condition (Fermat's rule):

$$\sigma^{(s)} \leftarrow \frac{\|\mathbf{y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2}{\sqrt{n}} \wedge \sigma_0 \ , \tag{14}$$

Algorithm 2 Reweighted $\text{MWE}_{0.5}$

Initialize weights $\mathbf{w}^{(s)} = \mathbb{1}$ for $s = 1 \dots S$
repeat
  Minimization: solve MWE_1 with the weighted ℓ_1 norms $\|\mathbf{w}^{(s)} \odot \mathbf{x}^{(s)}\|_1$
  Majorization: $\mathbf{w}_j^{(s)} = \frac{1}{2\sqrt{|\mathbf{x}_j^{(s)}|}}$ for all s, j
until convergence

Algorithm 3 MWE_1 algorithm

Input: σ_0, $\mu, \epsilon, \gamma, \lambda$ and cost matrix $\mathbf{M}$. data $(\mathbf{y}^{(s)})_s (\mathbf{L}^{(s)})_s$.
Output: MWE: $(\mathbf{x}^{(s)})$, minimizers of (13).
repeat
  for $s = 1$ **to** S **do**
    Update $\mathbf{x}^{(s)+}$ with proximal coordinate descent to solve (15).
    Update $\mathbf{x}^{(s)-}$ with proximal coordinate descent to solve (15).
    Update $\sigma^{(s)}$ with (14).
  end for
  Update left marginals $\mathbf{m}^{(1)+}, \dots, \mathbf{m}^{(S)+}$ and the barycenter $\bar{\boldsymbol{x}}^+$ with generalized Sinkhorn of Algorithm 1
  Update left marginals $\mathbf{m}^{(1)-} \dots, \mathbf{m}^{(S)-}$ and the barycenter $\bar{\boldsymbol{x}}^-$ with generalized Sinkhorn of Algorithm 1
until convergence

which also corresponds to the empirical estimator of the standard deviation when the constraint is not active. To update the remaining variables, we follow the same optimization procedure laid out in [32] and adapted to MWE in Algorithm 3. Briefly, let $\mathbf{m}^{(s)+} \stackrel{\text{def}}{=} \mathbf{P}^{(s)+}\mathbb{1}$ (resp. $\mathbf{m}^{(s)+} \stackrel{\text{def}}{=} \mathbf{P}^{(s)+}\mathbb{1}$), when minimizing with respect to one $\mathbf{x}^{(s)+}$ (resp. $\mathbf{x}^{(s)-}$), the resulting problem can be written (dropping the exponents for simplicity):

$$\min_{\mathbf{x}\in\mathbb{R}_+^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 + \frac{\mu\gamma}{S}(\langle \mathbf{x}, \mathbb{1}\rangle - \langle \log(\mathbf{x}), \mathbf{m}\rangle) + \lambda\sigma\|\mathbf{x}\|_1 \ , \tag{15}$$

which can be solved using proximal coordinate descent [15]. Note that the additional inference of a specific $\sigma^{(s)}$ for each subject allows to scale the Lasso penalty depending on their particular level of noise. The final update with respect to $((\mathbf{P}^{(s)+})_s, (\mathbf{P}^{(s)-})_s, \bar{\boldsymbol{x}}^+, \bar{\boldsymbol{x}}^-)$ can be cast as two Wasserstein barycenter problems, carried out using generalized Sinkhorn iterations [9]. Note that one does not need to compute the transport plans $P^{(s)}$ since inferring every source estimate $\mathbf{x}$ only requires the knowledge of the left marginal $\mathbf{m} = \mathbf{P}\mathbb{1}$ which does not require storing $\mathbf{P}$ in memory.

5. Experiments

5.1. Simulations with semi-real data

Benchmarks. As discussed in introduction, standard sparse source localization solvers are based on an ℓ_1 norm regularization, applied to the data of each subject independently. We use the independent Lasso estimator as a baseline. To illustrate the effect of reweighting separately, we also study the performance of a reweighted Lasso, i.e an independent regression with a $\ell_{0.5}$ penalty. Note that reweighted Lasso was coined iterative reweighted mixed-norm (irMxNE) in the context of M/EEG source imaging in [56]. We compare $\text{MWE}_{0.5}$ to the Group-Lasso estimator (6) [3, 69] which was proposed from multi-subject source imaging

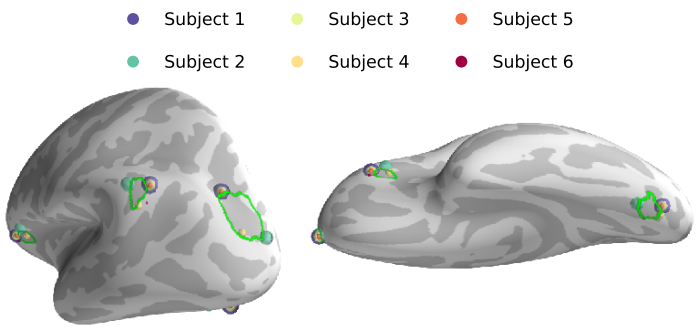

Fig. 3. Example of a simulated source configuration with 5 activations for $S = 6$ subjects - one activation per label. The 5 labels – highlighted within green borders – are taken from the aparc.a2009s FreeSurfer Destrieux parcellation [13]. Different radii are used to distinguish overlapping sources. Here, subjects 1, 3 and 5 share the exact same source locations.

to promote functional consistency across subjects [39]. We also evaluate the performance of a more flexible block sparse model where only a fraction of the source estimates are shared across all tasks: Dirty models [30]. In Dirty models source estimates are written as a sum of two parts which are penalized with different norms. One is common to all subjects (penalty ℓ_{21}) and one is specific for each subject (penalty ℓ_1). We also compare $\text{MWE}_{0.5}$ with MWE_1 to evaluate the benefits of non-convex penalties. For more details about the compared methods, we refer the reader to the appendix.

Hyperparameters of W. The parameters defining the Wasserstein distance $\widetilde{W}$ are ε (entropy regularization) and γ (marginal relaxation). Large values of ε accelerate the convergence of the Sinkhorn algorithm but induce an undesired blurring of the source estimates. Very Low values however lead to numerical instability. We set ε to 0.002 divided by the median of the ground metric M which provides a good trade-off between computation speed and sharpness of the barycenter. With the same reasoning, low values of γ allow for a "free" transport, thus the barycenter converges towards a blurred uniform distribution. We set γ to a lower bound $\gamma_M = -\frac{\max M}{2 \log 0.8} \approx 1$ that guarantees a minimal transport of mass using a strategy proposed in [32].

Simulation data and MEG/fMRI datasets. In our simulations, we use semi-real data, i.e. we simulate MEG data $\mathbf{y}$ with real leadfield matrices $\mathbf{L}$. To do so, we rely on the public Cam-CAN dataset [60]. We use the MRI scan of each subject to compute a source space and its associated leadfield comprising 2562 sources per hemisphere [22]. Keeping only MEG gradiometer channels, we have $n = 204$ observations per subject. To keep the simulation settings simple, we restrict all leadfields to the left hemisphere. We thus have $S = 32$ leadfields with $p = 2562$. We simulate an inverse solution $\mathbf{x}^s$ with 5 sources (5-sparse vector) by randomly selecting one source per label (a.k.a. region of interest) among 5 pre-defined labels using the *aparc.a2009s* parcellation of the Destrieux atlas [13]. To model functional consistency, 50% of the subjects share sources at the same locations, the remaining 50% have sources randomly generated in the same labels (see Figure 3 for an example). Their

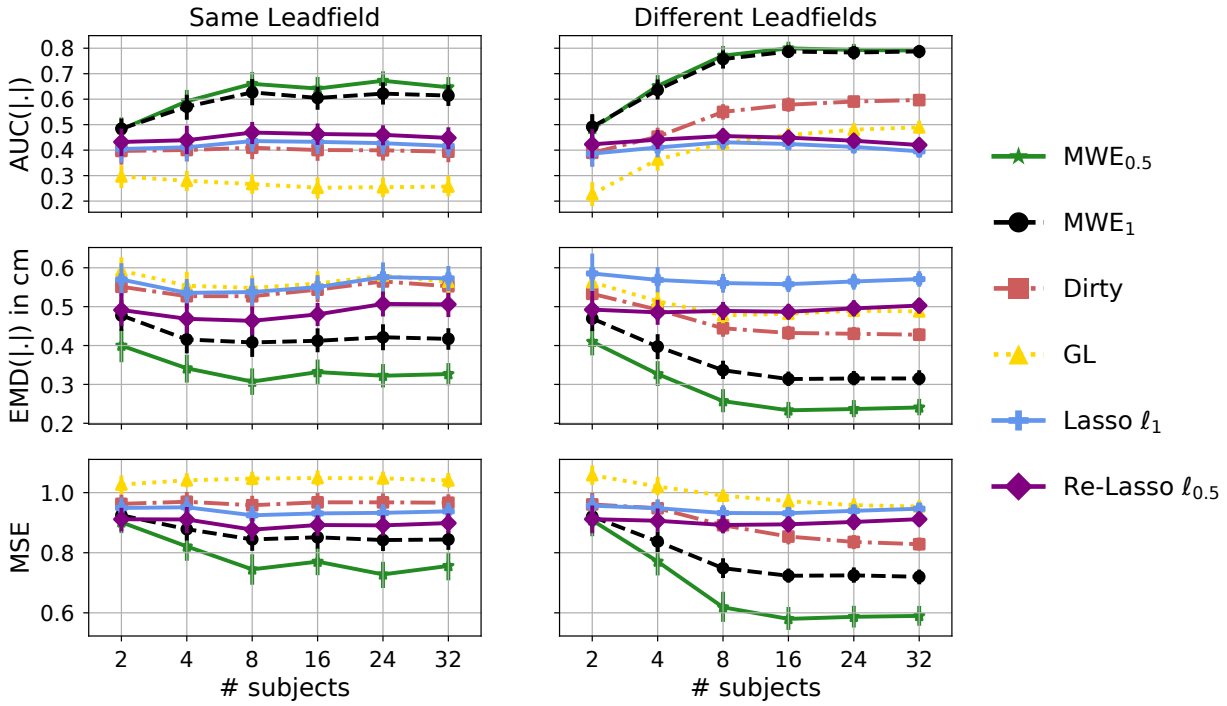

Fig. 4. Performance of different models over 30 trials in terms of AUC, EMD and MSE using the same leadfield for all subjects (randomly selected in each trial) (**left**) and different leadfields (**right**) computed using Cam-CAN dataset with 5 simulated sources.

amplitudes are taken uniformly between 20 and 30 nAm. Their sign is taken at random with a Bernoulli distribution (0.5) for each label (Hence all subjects share the same polarity of currents in a given label). We simulate $\mathbf{y}$ using the forward model with a covariance matrix σI_n. We set σ so as to have an average signal-to-noise ratio across subjects equal to 4 ($\text{SNR} \overset{\text{def}}{=} \sum_{s=1}^{S} \frac{\|\mathbf{L}^{(s)}\mathbf{x}^{(s)}\|}{S\sigma}$).

We evaluate the performance of all models knowing the ground truth by comparing the best estimates on a grid of hyperparameters in terms of three metrics: the mean squared error (MSE) to quantify accuracy in amplitude estimation, AUC and a generalized Earth mover distance (EMD) to assess supports estimation. We use the PR-AUC (Area under the curve Precision-recall) computed between the absolute values of the coefficients and the true supports. Similarly, the EMD is computed between normalized absolute values of sources. Since $\mathbf{M}$ is expressed in millimeters, EMD can be seen as an expectation of the geodesic distance between the truth and the source estimates. For a better intuitive interpretation of the EMD, we compute the EMD per source i.e we divide it by 5. The mean across subjects is reported for all metrics.

Simulation results. We vary the number of subjects under two conditions: (1) using the same leadfield for all subjects, as one would do with a template head model, (2) using specific leadfield operators of each subject. Each model is fitted on a grid of hyperparameters and the best AUC/MSE/EMD scores are reported. We perform 30 different trials (with different

true activations and noise, different common leadfield for condition (1)) and report the mean within a 95% confidence interval in Figure 4.

Various observations can be made. First by comparing conditions (1) and (2), one can notice the benefit of using different leadfield operators across subjects. This gain in performance concerns all multi-task models, especially OT based models MWE_q. We argue that this improvement is the consequence of the different folding patterns of the cortex across subjects. Indeed, these folding differences lead to different dipole orientations of the same source across subjects, thereby increasing the chances of an accurate localization. Second, note that the Group Lasso [39] performs poorly – even compared to independent Lasso – which is expected since simulated sources are not perfectly overlapping for all subjects. Indeed, in the simple case of 2 subjects, on can show that if the fraction of overlapping sources is less than 2/3, Group Lasso performs worse than independent Lasso [47]. Our experiments confirm this theoretical result. MWE_q however benefits from the presence of more subjects by leveraging spatial proximity. The mean AUC increases from 0.4 (Lasso) to 0.8. The average error EMD distance is reduced from 6 mm (Lasso) to nearly 2 mm. Finally, even if both MWE_q models show a similar AUC score, the proposed reweighting allows $\text{MWE}_{0.5}$ to outperform MWE_1 by a significant margin in terms of amplitude estimation (MSE). Finally, by inducing more sparsity, the $\ell_{0.5}$ norm of $\text{MWE}_{0.5}$ reduces the number of false positives which are located far from the true sources, thereby reducing the EMD distance by 1mm compared to MWE_1.

5.2. Experiments on MEG data

Datasets description. The different strategies were evaluated on two publicly available MEG datasets: DS117 [66] and Cam-CAN [60]. DS117 provides MRI, MEG, EEG and fMRI data of 16 healthy subjects to whom were presented images of famous, unfamiliar and scrambled faces. The fusiform face area (FFA) which specializes in facial recognition activates around 170ms after stimulus [29, 34]. We pick the time point in the contrast response *famous* vs *scrambled* with the peak response for each subject within the interval 150-200ms after stimulus. Similarly, Cam-CAN provides MEG, EEG and MRI data of around 650 healthy subjects with several types of tasks. We select the youngest 32 subjects (aged between 18 years and 29 years) and use their MEG recordings to study the auditory N100 response. We average the responses of 3 stimuli: 300Hz, 600Hz and 1200Hz with a total of 60 trials. We pick the time point with the peak response within 80-120 ms after stimulus. For both datasets, the leadfield operator of each subject was obtained from their T1 MRI scan using a cortically constrained source space formed by about 2500 candidate dipoles per hemisphere.

Model selection. For all lasso-type models, there exists $\lambda_{\max}$ such that for $\lambda \geq \lambda_{\max}$ the inverse solution is 0 everywhere. For instance, with ℓ_1 and $\ell_{0.5}$ we have $\lambda_{\max} = \frac{\|\mathbf{L}^\top \mathbf{y}\|_\infty}{n}$ [54]. This allows to set λ in a relative scale between 0 and 1, making this choice less sensitive to the data. In practice, one can pick a certain value in [0, 1] based on the number of active sources, which is the heuristic used in the following experiments with real data. Even though the choice of $\lambda_{\max}$ does not theoretically guarantee null source estimates with $\text{MWE}_{0.5}$, we observe experimentally that reweighting and the OT regularizer promote even more sparsity

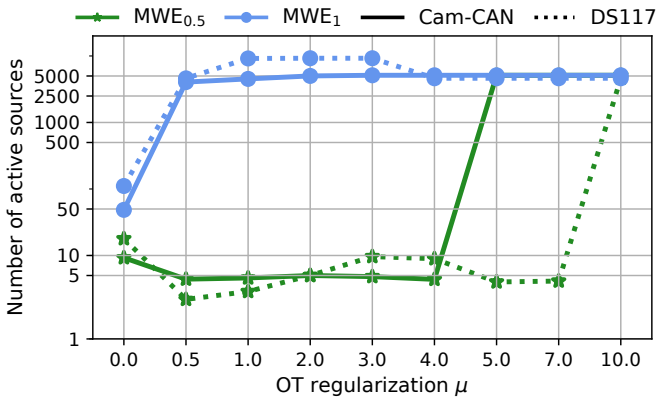

Fig. 5. Number of active sources for MWE models with $\lambda = 30\%$. The mean is reported across all subjects. With reweighted MWE, a similar phase transition occurs for both datasets after a certain $\mu_{\max}$.

with a lower λ compared to Lasso models. We use the same relative scaling to set λ for $\text{MWE}_{0.5}$. The OT regularization parameter μ controls the level of consistency across subjects. Figure 5 shows that for the reweighted $\text{MWE}_{0.5}$, there exists a phase transition at a certain value $\mu_{\max}$, after which the source estimates lose all sparsity and cover the entire cortical mantle uniformly. MWE_1 however shrinks the source estimates towards 0 but fails to produce sparse solutions. In practice, based on the complexity of the topographic maps of the MEG data, we select λ and μ that lead to – on average – a 2-sparse solution with Cam-CAN ($\lambda = 30\%, \mu = 3$) and a 6-sparse solution with DS117 ($\lambda = 20\%, \mu = 0.5$).

MWE for population imaging. The standard approach to obtain the source estimates from a group of subjects is to average the estimates obtained independently for each subject. Euclidean averaging however induces undesired blurring and sparsity is lost even when the individual solutions are sparse. Figure 7 shows that $\text{MWE}_{0.5}$ prevents that from happening. Moreover, the latent variable $\bar{\boldsymbol{x}}$ of $\text{MWE}_{0.5}$ is sharper and more informative at a population level. To compare with single-subject solvers, we compute MCE and reweighted MCE solutions by selecting independently for each subject a λ such that the solution is 2-sparse (resp. 6-sparse) for Cam-CAN (resp. DS117). For dSPM, we use the default hyperparameter value $1/\text{SNR}^2$ with $\text{SNR} = 3$. The green borders highlight regions of interest. For Cam-CAN, we use the *neurosynth* [68] label corresponding to the *auditory cortex* thresholded at 15 and projected on the surface of the temporal lobe. For DS117, we rely on the *aparc a2009s* segmentation to show both the fusiform gyrus and the primary visual cortex V1. With Cam-CAN, the Euclidean average of the obtained minimum Wasserstein estimates is focal, located right in the auditory cortex. However, The average Lasso and dSPM estimates are dispersed around the auditory cortex with a substantial blurring due to averaging. The visual task of DS117 appears to be the most challenging for several reasons which explain the low amplitude sources. These reasons are discussed in detail in section 6.

Comparison with fMRI. The EEG/MEG inverse problem has an infinite number of solutions. We proposed to regularize it in two ways: (1) at a subject level by favoring focal sources; (2)

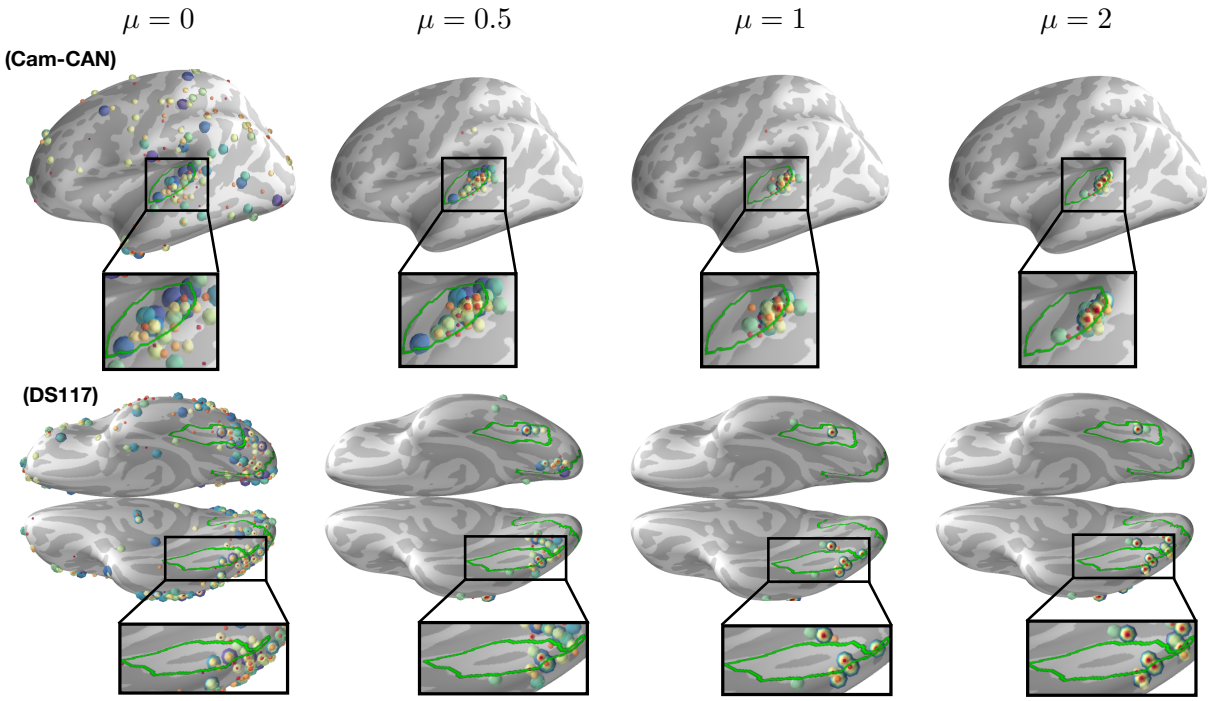

Fig. 6. Support of source estimates of $\text{MWE}_{0.5}$ recovered in the auditory task of Cam-CAN with 32 subjects (top) and the visual task of DS117 with 16 subjects (bottom). Each color corresponds to a subject. Different radii are displayed for a better distinction of sources. Increasing μ with $\mu < \mu_{\max}$ promotes functional consistency across subjects. Top: Cam-CAN dataset ($\lambda = 30\%$). Bottom: DS117 dataset ($\lambda = 20\%$).

at a population level by promoting spatial proximity between activation foci. However, one could argue that $\text{MWE}_{0.5}$ promotes consistency at the expense of proper fitting of individual data. To address this concern we compute the standardized fMRI Z-score of the conditions *famous vs scrambled faces*. We compare minimum current estimates (MCE or Lasso) [63], reweighted MCE, $\text{MWE}_{0.5}$ and fMRI by computing for each subject the mean geodesic distance between the mode of the neural activation map of each subject and the vertices of the Fusiform-gyrus (FFG) as well as the primary visual cortex (V1). Figure 8 shows that the distribution of MWE geodesics is closer to that of fMRI z-maps. By promoting functional similarity, MWE disregards the spurious activation that are far from the regions of interest. Moreover, one can notice that some 6-sparse MCE models cancel out all sources in the left hemisphere (subjects with a geodesic equal to $+\infty$).

6. Discussion

The M/EEG source imaging problem is a notoriously hard inverse problem, in particular when the underlying neural activity is distributed over different coactive brain regions. To tackle this problem, this work proposes to jointly localize sources for a population of subjects by casting the estimation as a multi-task regression problem.

Embracing this formulation of multi-task regression, this work develops three key ideas. First it proposes to use non-linear registration to obtain subject specific leadfield matrices

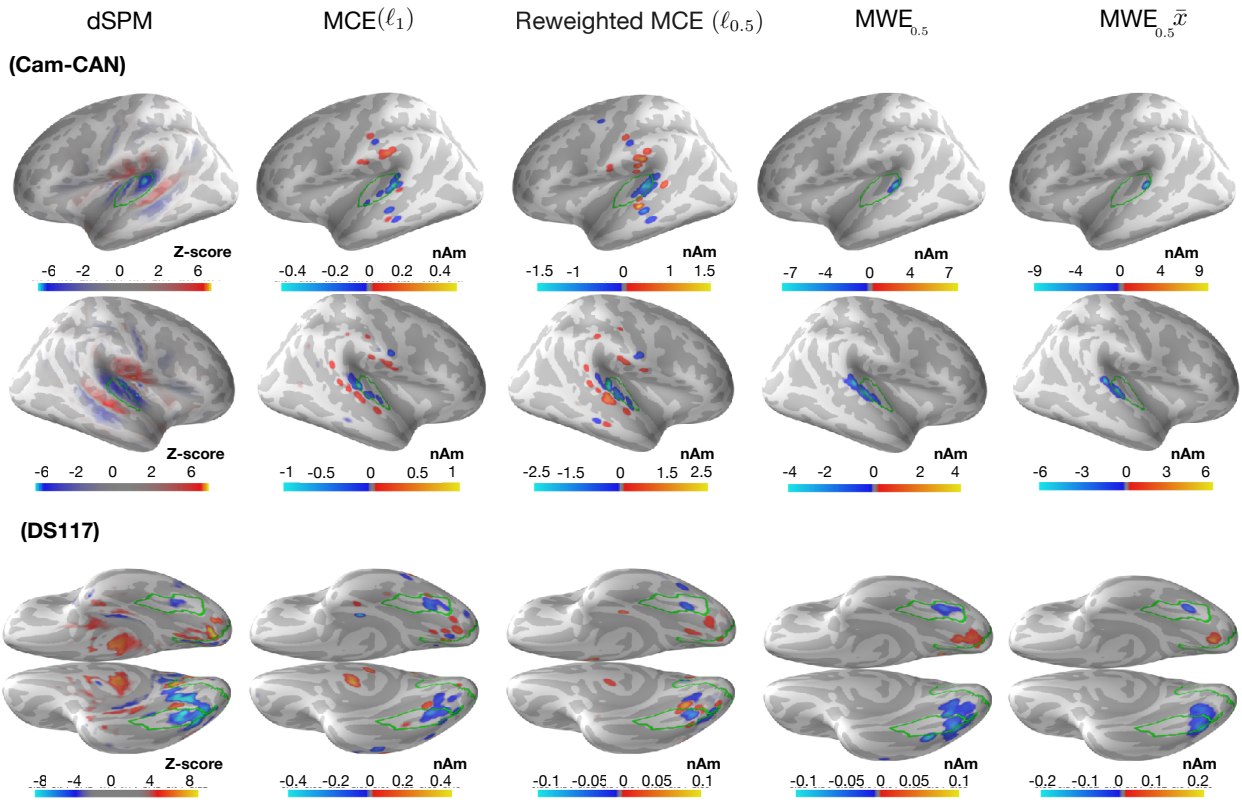

Fig. 7. Average source estimates of different solvers. **Top:** Cam-CAN dataset. **Bottom:** DS117 dataset. $\text{MWE}_{0.5}$ $\bar{\boldsymbol{x}}$ is the latent variable inferred in the $\text{MWE}_{0.5}$ model, corresponding to a Wasserstein barycenter of the $\text{MWE}_{0.5}$ source estimates. $\text{MWE}_{0.5}$ reduces blurring by promoting functional consistency.

that are spatially aligned. Second it copes with the issue of inter-subject spatial variability of functional activations using Wasserstein metrics and optimal transport theory. Finally, it makes use of advanced techniques from the inverse problem literature using sparsity promoting priors. This allows to model variations of recordings in terms of noise levels using concomitant estimation of sources and noise amplitudes, and it uses ℓ_q quasi-norms with $q < 1$ to obtain more accurate source amplitudes.

The classic pipeline of a M-EEG group source imaging study is to perform source localization independently across subjects using inverse solvers such as MNE, MCE, sLORETA, dSPM or MxNE. The group-level analysis is then carried out as a post-processing step by averaging the source estimates of each subject or by aggregating Z-scores in a multiple tests comparison [59]. This is usually done thanks to a non-linear registration and by averaging of the estimates after mapping them to the same brain template. In this work, a different approach based on multi-task regression is proposed. The non-linear registration is used to compute leadfield matrices that are spatially aligned. A source space formed by candidate dipoles are defined on the template brain geometry and this source space is warped to individual anatomies for which Maxwell equations are solved numerically. By doing so, we demonstrate improvements in terms of source localization accuracy. This is significant

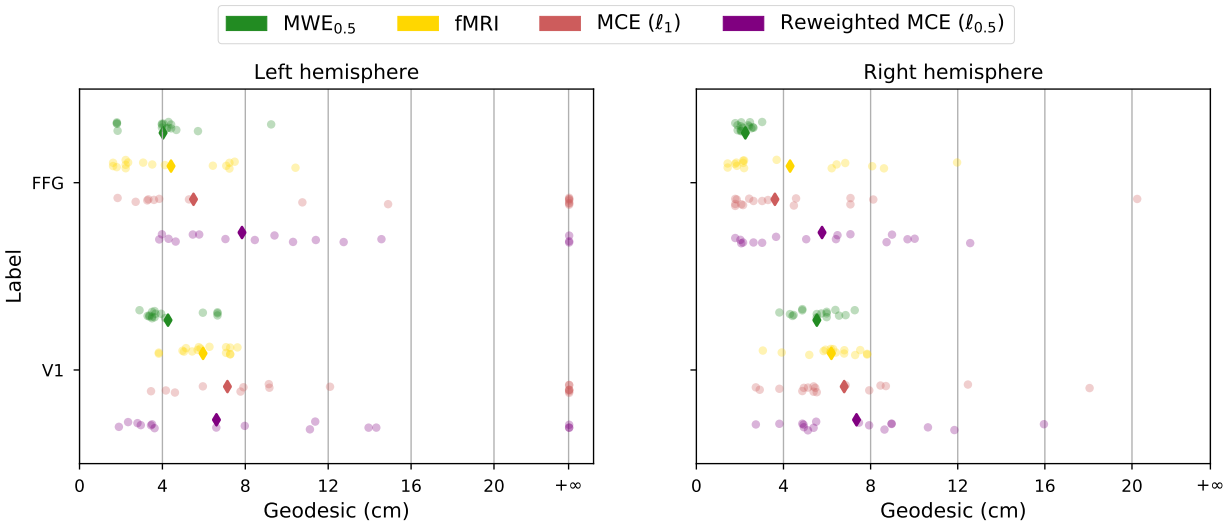

Fig. 8. Mean geodesic distance between the mode of the M/EEG derived neural activation map and the vertices of the labels FFA and V1. Each dot represents one of the 16 subjects. For some subjects, MCE / reweighted MCE produce 6-sparse solutions entirely in the right hemisphere, to which the goedesic $+\infty$ is assigned.

evidence that anatomical variability can be more a blessing than a curse for group level M/EEG source imaging.

This statement is actually inline with the work of Larson et al. [38], who suggested that anatomical differences between subjects can improve the accuracy of the averaged source estimates by emphasizing common sources across subjects. Our simulations confirm this hypothesis not only for averaged estimates but also for individual ones. Indeed, all the multi-task models studied in our simulations improve with more subjects only if the used leadfield operators are different. This striking result suggests that using the same head model in M/EEG group studies for different subjects potentially causes a significant loss of information. One possible explanation of why anatomical differences help is that anatomical variability combined with functional similarities lead to non-redundant information across subjects. Take the example of a shared source across subjects. Different folding patterns of the cortical mantle would lead to different (normal) orientations of the current dipole. Since the relative position of the sensors is not changed, the leadfields – having different sensitivity maps – would generate measurements with more information, i.e higher rank. On the contrary, when using the same leadfield for all subjects, an exact same source would lead to similar M/EEG measurements. Quantitatively, our simulations with semi-real data show that multi-subject inverse solvers improve the localization error by almost 4mm per source with different leadfields and only 1 mm when the same leadfield is used.

By pooling together data from multiple subjects one can increase the number of measurements, hence make the problem less ill-posed. Yet, this cannot be done without taking into consideration differences between subjects, especially the spatial variability in activation patterns. To cope with this issue when averaging brain patterns both in M/EEG and fMRI,

Wasserstein distances have proven efficient [23]. Through this work, we explained how they could be included directly in the inverse solver. Thanks to their ability to model spatial proximity between source estimates, the MWE model allows to promote functional similarities across subjects using the geometry of the cortical mantle. Fortunately, the computation of the Wasserstein barycenter does not lead to a computational bottleneck. In our experiments, 40% to 60% of time is spent on optimal transport versus proximal coordinate descent. Thanks to careful optimization procedures based on Sinkhorn iterations and block coordinate descent algorithms, the model proposed here runs in a few minutes on empirical M/EEG datasets.

Beyond the use of Wasserstein metrics to cope with spatial misalignments, the proposed MWE_q model brings in two important ingredients from the statistics literature employing sparsity promoting regularizations: concomitant estimation and convex reweighted schemes. By using concomitant estimation, the MWE_q model can cope with the different noise levels and signal-to-noise ratios for the different subjects. This is particularly critical to have the number of hyperparameters of the model that is fixed and does not scale with the number of subjects. In theory, for source imaging with a solver such as dSPM or sLORETA, that is applied independently for all subjects, the regularization parameters could be tuned for each dataset. The MWE_q model has a list of regularization parameters that does not depend on the number of subjects. Besides, results from Figures 4 and 5 demonstrate the benefit of $\text{MWE}_{0.5}$ vs. MWE_1. Employing a more aggressive sparsity promoting regularization improves in particular the source amplitude estimation as shown by the MSE metric. Also it greatly simplifies the setting of the regularization parameter μ as solutions become suddenly much less sensitive to this choice of parameter.

From a more neuroscientific perspective, the model presented here has potentially interesting consequences. Results on Cam-CAN demonstrate that the barycenters obtained with $\text{MWE}_{0.5}$ have a higher spatial specificity. As seen in Figure 7, activation foci in $\bar{\boldsymbol{x}}$ are well limited to primary auditory cortices while solvers that are not based on a group-level multi-task regression model lead to spurious activations next to secondary somatosensory cortices and on middle temporal gyrus. On DS117 dataset, the cognitive task performed by the subjects is more advanced, complicating the discussion of the results in terms of localization. Yet, the availability of the fMRI data allows for a quantification of the activation foci between MEG and fMRI. While it is often repeated that fMRI and M/EEG sources are different, and thus brain activation maps obtained by these different modalities should not necessarily match, our findings demonstrate that the proposed method reduces the gap between MEG source imaging and fMRI.

Appendix

For the sake of clarity, we provide in this appendix some technical background on all the models discussed in this work.

Adaptive-Lasso

In its general framework, the adaptive Lasso replaces the ℓ_1 penalty in (4) with a separable non-convex function. Let $g : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave differentiable function on $\mathbb{R}_{++}$. The adaptive Lasso is defined as:

$$\mathbf{x}^\star = \underset{\mathbf{x}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{2n}\|\mathbf{y}-\mathbf{L}\mathbf{x}\|_2^2+\lambda\sum_{j=1}^{p} g(|\mathbf{x}_j|)\ . \tag{16}$$

Gasso et al. [18] showed that problem (16) can be solved iteratively using nested weighted Lasso problems illustrated in algorithm 4. Taking $g : x \to \sqrt{x}$ leads to the $\ell_{0.5}$ reweighting presented in section 4.

Group Lasso

The Group Lasso [39, 69] solves several related regression jointly by enforcing a form of structured sparsity. In the context of source imaging, this structured sparsity corresponds to a strict consensus across subjects to decide whether a source is active or not. Formally, the group lasso solves:

$$\min_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(S)}\in\mathbb{R}^p}\ \frac{1}{2n}\sum_{s=1}^{S}\|\mathbf{y}^{(s)}-\mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2+\sum_{j=1}^{p}\sqrt{\sum_{s=1}^{S}{\mathbf{x}_j^{(s)}}^2}\ . \tag{17}$$

The double sum penalty can be seen as an ℓ_1 penalty applied to a vector of ℓ_2 norms taken across subjects: only some ℓ_2 norms are non-zero. Therefore, source are cancelled out for all subjects or for none of them.

Dirty models

The assumption of identical sources for all subjects is clearly not realistic, Dirty models [30] relax this assumption by decomposing the source vector of each subject s into two parts: $\mathbf{x}^{(s)} = \mathbf{x}_c^{(s)} + \mathbf{x}_s^{(s)}$, where the support of $\mathbf{x}_c^{(s)}$ is common to all subjects and $\mathbf{x}_s^{(s)}$ is specific to each one. The regularization then writes:

$$J_{\text{Dirty}}(\mathbf{x}_c,\mathbf{x}_s) = \mu\|(\mathbf{x}_c^{(1)},\ldots,\mathbf{x}_c^{(s)})\|_{21}+\lambda\sum_{s=1}^{S}\|\mathbf{x}_s^{(s)}\|_1\ .$$

Dirty models hence aim at solving the optimization problem:

$$\min_{\mathbf{x}_c,\mathbf{x}_s}\frac{1}{2n}\sum_{s=1}^{S}\|\mathbf{L}^{(s)}(\mathbf{x}_c^{(s)}+\mathbf{x}_s^{(s)})-\mathbf{y}^{(s)}\|_2^2+J_{\text{Dirty}}(\mathbf{x}_c,\mathbf{x}_s)\ .$$

Algorithm 4 Adaptive Lasso reweighting

Initialize weights $\mathbf{w}^{(s)} = \mathbb{1}$ for $s = 1 \ldots S$
repeat
  Minimization: solve Lasso with the weighted ℓ_1 norms $\|\mathbf{w}^{(s)} \odot \mathbf{x}^{(s)}\|_1$
  Majorization: $\mathbf{w}_j^{(s)} = g'(\mathbf{x}_j^{(s)})$ for all s, j
until convergence

When $\mathbf{x}_s = \mathbf{0}$ (resp. $\mathbf{x}_c = \mathbf{0}$) one falls back to a Group Lasso (resp. independent Lasso). Indeed, the ℓ_{21} norm forces the $\mathbf{x}_c$ to share the same active locations across subjects. The advantage of Dirty models over the Group Lasso is that it is agnostic with respect to the degree of similarities across subjects.

Acknowledgements

This work was funded by the ERC Starting Grant SLAB ERC-YStG-676943 and a *chaire d'excellence de l'IDEX Paris Saclay.*

References

[1] Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM*, 43(2):904–924.

[2] Ahlfors, S. P., Ilmoniemi, R. J., and Hämäläinen, M. S. (1992). Estimates of visually evoked cortical currents. *Electroencephalography and Clinical Neurophysiology*, 82(3):225–236.

[3] Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Advances in Neural Information Processing Systems*.

[4] Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, 20:327 EP –.

[5] Baillet, S., Mosher, J. C., and Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30.

[6] Becker, H., Albera, L., Comon, P., Gribonval, R., Wendling, F., and Merlet, I. (2015). Brain-Source Imaging: From sparse to tensor models. *IEEE Signal Processing Magazine*, 32(6):100–112.

[7] Benamou, J., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman Projections For Regularized Transportation Problems. *Society for Industrial and Applied Mathematics*.

[8] Candès, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905.

[9] Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2017). Scaling Algorithms for Unbalanced Transport Problems. arXiv Preprint 1607.05816.

[10] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300.

[11] Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping. *Neuron*, 26(1):55–67.

[12] Deslauriers-Gauthier, S., Lina, J.-M., Butler, R., Bernier, P.-M., Whittingstall, K., Deriche, R., and Descoteaux, M. (2017). Inference and Visualization of Information Flow in the Visual Pathway using dMRI and EEG. In *MICCAI 2017 Medical Image Computing and Computer Assisted Intervention*, Québec, Canada.

[13] Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1 – 15.

[14] Engemann, D. A. and Gramfort, A. (2015). Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage*, 108:328 – 342.

[15] Fercoq, O. and Richtárik, P. (2015). Accelerated, parallel and proximal coordinate descent. *SIAM Journal on Optimization*, 25:1997–2023.

[16] Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9:195 – 207. Mathematics in Brain Imaging.

[17] Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. A. (2015). Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems (NIPS) 28*.

[18] Gasso, G., Rakotomamonjy, A., and Canu, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698.

[19] Gorodnitsky, I. F. and Rao, B. D. (1997). Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616.

[20] Gramfort, A., Kowalski, M., and Hämäläinen, M. (2012). Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in Medicine and Biology*, 57(7):1937–1961.

[21] Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013a). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267.

[22] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M. (2013b). MNE software for processing MEG and EEG data. *NeuroImage*, 86.

[23] Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast optimal transport averaging of neuroimaging data. In *Proceedings of the Information Processing in Medical Imaging conference*.

[24] Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M., and Kowalski, M. (2013c). Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70(0):410 – 422.

[25] Hämäläinen, M. S. and Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, 32(1):35–42.

[26] Hämäläinen, M. S. and Sarvas, J. (1987). Feasibility of the homogeneous head model in the interpretation of neuromagnetic fields. *Physics in Medicine and Biology*, 32(1):91.

[27] Hari, R. and Puce, A. (2017). *MEG–EEG Primer*. DOI 10.1093/med/9780190497774.003.0001.

[28] Haufe, S., Nikulin, V. V., Ziehe, A., Müller, K.-R., and Nolte, G. (2008). Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2):726–38.

[29] Henson, R. N., Wakeman, D. G., Litvak, V., and Friston, K. J. (2011). A parametric empirical bayesian framework for the EEG/MEG inverse problem: Generative models for multi-subject and multi-modal integration. *Frontiers in human neuroscience*, 5:76; 76–76.

[30] Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. (2010). A Dirty Model for Multi-task Learning. *Advances in Neural Information Processing Systems*.

[31] Janati, H., Bazeille, T., Thirion, B., Cuturi, M., and Gramfort, A. (2019a). Group level EEG/MEG source imaging via Optimal Transport: minimum Wasserstein estimates. In *Proceedings of the Information Processing in Medical Imaging conference*.

[32] Janati, H., Cuturi, M., and Gramfort, A. (2019b). Wasserstein regularization for sparse multi-task regression. In *Proceedings of Machine Learning Research*, volume 89. PMLR.

[33] Kantorovic, L. and Rubinshteïn, G. S. (1957). *On a functional space and certain extremum problems*, volume 115. Doklady Akademii Nauk SSSR.

[34] Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311.

[35] Knopp, P. and Sinkhorn, R. (1967). Concerning nonnegative matrices and doubly stochastic matrices. . *Pacific Journal of Mathematics*, 1(2):343–348.

[36] Kozunov, V. V. and Ossadtchi, A. (2015). Gala: group analysis leads to accuracy, a novel approach for solving the inverse problem in exploratory analysis of group MEG recordings. *Frontiers in Neuroscience*, 9:107.

[37] Kybic, J., Clerc, M., Abboud, T., Faugeras, O., Keriven, R., and Papadopoulo, T. (2005). A common formalism for the integral formulations of the forward EEG problem. *IEEE Transactions on Medical Imaging*, 24(1):12–28.

[38] Larson, E., Maddox, R. K., and Lee, A. K. C. (2014). Improving spatial localization in MEG inverse imaging by leveraging intersubject anatomical differences. *Frontiers in Neuroscience*, 8:330.

[39] Lim, M., Ales, J., Cottereau, B. M., Hastie, T., and Norcia, A. M. (2017). Sparse eeg/meg source estimation via a group lasso. *PLOS*.

[40] Lin, F.-H., Witzel, T., Ahlfors, S. P., Stufflebeam, S. M., Belliveau, J. W., and Hämäläinen, M. S. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage*, 31(1):160–171.

[41] Litvak, V. and Friston, K. (2008). Electromagnetic source reconstruction for group studies. *NeuroImage*,

42(4):1490 – 1498.

[42] Mainini, E. (2012). A description of transport cost for signed measures. *Journal of Mathematical Sciences*, 181(6):837–855.

[43] Massias, M., Fercoq, O., Gramfort, A., and Salmon, J. (2018). Generalized concomitant multi-task lasso for sparse multimodal regression. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of Machine Learning Research*, volume 84, pages 998–1007. PMLR.

[44] Michel, C. M., Murray, M. M., Lantz, G., Gonzalez, S., Spinelli, L., and de Peralta, R. G. (2004). EEG source imaging. *Clinical Neurophysiology*, 115(10):2195–2222.

[45] Mosher, J., Leahy, R., and Lewis, P. (1999). EEG and MEG: Forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46(3):245–259.

[46] Ndiaye, E., Fercoq, O., Gramfort, A., Leclère, V., and Salmon, J. (2017). Efficient smoothed concomitant Lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 904(1):012006.

[47] Negahban, S. and Wainwright, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of l1,∞ regularization. *Advances in Neural Information Processing Systems*.

[48] Okada, Y. (1993). Empirical bases for constraints in current-imaging algorithms. *Brain Topography*, 5:373–377.

[49] Ou, W., Hämäläinen, M., and Golland, P. (2009). A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946.

[50] Ou, W., Nummenmaa, A., Ahveninen, J., Belliveau, J. W., Hämäläinen, M. S., and Golland, P. (2010). Multimodal functional imaging using fMRI-informed regional EEG/MEG source estimation. *NeuroImage*, 52(1):97 – 108.

[51] Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72.

[52] Pascual-Marqui, R. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol*, 24:D:5–12.

[53] Profeta, A. and Sturm, K.-T. (2018). Heat flow with dirichlet boundary conditions via optimal transport and gluing of metric measure spaces. arXiv Preprint 1809.00936.

[54] Rakotomamonjy, A., Gasso, G., and Salmon, J. (2019). Screening rules for Lasso with non-convex sparse regularizers. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5341–5350, Long Beach, California, USA. PMLR.

[55] Sato, M., Yamashita, O., Sato, M.-A., and Miyawaki, Y. (2018). Information spreading by a combination of meg source estimation and multivariate pattern classification. *PloS one*, 13(6):e0198806–e0198806.

[56] Strohmeier, D., Bekhti, Y., Haueisen, J., and Gramfort, A. (2016). The iterative reweighted mixed-norm estimate for spatio-temporal MEG/EEG source reconstruction. *IEEE Transactions on Medical Imaging*, 35(10):2218–2228.

[57] Strohmeier, D., Gramfort, A., and Haueisen, J. (2015). Meg/eeg source imaging with a non-convex penalty in the time-frequency domain. pages 21–24.

[58] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99:879–898.

[59] Takeda, Y., Suzuki, K., Kawato, M., and Yamashita, O. (2019). MEG Source Imaging and Group Analysis Using VBMEG. *Frontiers in Neuroscience*, 13:241.

[60] Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., Tyler, L. K., Cam-CAN, and Henson, R. N. (2017). The cambridge centre for ageing and neuroscience (Cam-CAN) data repository: Structural and functionalMRI , meg, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262 – 269. Data Sharing Part II.

[61] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

[62] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

[63] Uutela, K., Hämäläinen, M. S., and Somersalo, E. (1999). Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage*, 10(2):173–180.

[64] Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information Processing in Medical Imaging*, volume 6801, pages 562–573. Springer.

[65] Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer Verlag.

[66] Wakeman, D. and Henson, R. (2015). A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2(150001).

[67] Wipf, D. and Nagarajan, S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966.

[68] Yarkoni, T. (2014). Neurosynth core tools v0.3.1.

[69] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67.

[70] Zhongming Liu, Lei Ding, and Bin He (2006). Integration of EEG/MEG with MRI and fMRI. *IEEE Engineering in Medicine and Biology Magazine*, 25(4):46–53.